Video Stream Retrieval of Unseen Queries using Semantic Memory

Spencer Cappallo, Thomas Mensink, Cees G.M. Snoek
University of Amsterdam
Retrieval for live streaming video?

Many live video services (Periscope, Twitch, Meerkat)

Limited to visual content

→ Challenging, motivated no-example problem
Motivation: Two Tasks

Instantaneous Retrieval
“Which streams have dancing right now?”

Continuous Retrieval
“Just keep showing dancing”
Stream Retrieval

- No information from future

- Wide range of queries

- Relevant to present content
Stream Retrieval

- No information from future
- Wide range of queries
- Relevant to present content
People search for everything...

“UFO Invasion Belarus”

→ Never happened before, no classifiers

→ Adapt Zero-shot Classification technique
Concepts to Queries

Semantic embedding relates query to pre-trained concepts (Norouzi, 2013)

Query: “Carpentry”
Related Work: Zero-shot on Videos

Image concepts used to predict video actions (Jain, 2015)

- No temporal evaluation
- Whole video available
- Limited set of actions
Stream Retrieval

- No information from future
- Wide range of queries
- Relevant to present content
Memory for Stream Retrieval

Representation must reflect what is happening *now*

$\rightarrow \textbf{Memory}$ to prioritize recent information
Mean and Max Memory Pooling

Mean or Max Pooling over memory window
Memory Pooling Drawbacks

- Discards all information past memory
- $m$ frames per concept per stream
- Arbitrary selection of top concepts
Memory Welling

Instead of temporal pooling, well fills and drains over time...
Memory Welling

Well defined by:

\[ w(x_t) = \max \left( \frac{m-1}{m} w(x_{t-1}) + \frac{1}{m} x_t - \beta, \ 0 \right) \]

- \( m \) is memory parameter
- \( \beta \) is a constant “leakiness” term
  - Sparsity
  - Concept reliability

→ Emphasizes reliable, recent information
## Pooling vs Welling

<table>
<thead>
<tr>
<th>Memory Pooling</th>
<th>Memory Welling</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Only uses $m$ frames of information</td>
<td>+ No hard memory cut-off</td>
</tr>
<tr>
<td>- $m$ frames per feature per stream</td>
<td>+ Only current state stored</td>
</tr>
<tr>
<td>- Arbitrary selection of top concepts</td>
<td>+ Sparsity enforced implicitly</td>
</tr>
</tbody>
</table>

**Memory Welling addresses limitations of Pooling**
How to experiment?

No video stream retrieval data set...
...but existing datasets can be adapted

- Evaluate videos “concurrently”
- Algorithm sees only present and past frames
- Validation & Test split of disjoint classes
Experimental Setup Details

Datasets

- **ActivityNet**, annotation has temporal extent
- **AN-L**, concatenated videos, simulates changing content

Features

- 13k ImageNet Concepts, GoogLeNet
- 2 frames / second
Task Evaluation

- **Instantaneous Retrieval**
  Which streams are relevant *right now*?

- **Continuous Retrieval**
  Keep showing me relevant content
Instantaneous Retrieval: Evaluation

Which videos are relevant now?

Measure with mean AP across time:

\[
\frac{1}{\sum_{t} y_{t}} \sum_{t} \text{AP}_{t} y_{t}
\]
Instantaneous Retrieval - ActivityNet Results

→ Memory welling outperforms

→ Shines with temporal annotation
Analysis - Memory Parameter

Performance vs. Memory Size

- Max Pooling
- Mean Pooling
- Memory Welling

Best memory equivalent to ~12 seconds
Task Evaluation

- **Instantaneous Retrieval**
  Which streams are relevant *right now*?

- **Continuous Retrieval**
  Keep showing me relevant content
Task 2: Continuous Retrieval

“Keep showing me relevant content”
→ e.g., a user wants to watch dancing for thirty minutes
Task 2: Continuous Retrieval Evaluation

- Reward relevant stream
- Penalize needless switches
- Temporal consistency

Our metric: \[
\frac{z_+ + r_+}{\sum_t y^t}
\]

\(z_+\) counts ‘zaps’ from irrelevant to relevant stream
\(r_+\) counts remaining on relevant stream
Continuous Retrieval - Results AN-L

→ Approach copes with changing content
Conclusion

➢ We explore new problem of no-example stream retrieval
➢ We outline need and approach for temporal evaluation
➢ We introduce and test three memory-based approaches

Thank you!
Contact: cappallo@uva.nl
Analysis - Annotation Type

Frame concepts are objects

→ Strongest on objects

→ Temporal concepts?
Stream Retrieval

No-example retrieval approach adapted from zero-shot classification

A semantic embedding relates query terms to visual concepts:

\[ \text{score}(q, x_t) = s(q)^T \phi(x_t) \]

Where \( x \) limited to present and past frames

\( \Phi \) creates sparsity
Stream Retrieval

Live Video Streams

Deep Network

Semantic Embedding Space

Input Query

"ufo invasion Belarus"

1. 
2. 
3. 

Retrieved Streams
## Instantaneous Retrieval - Results

<table>
<thead>
<tr>
<th></th>
<th>AN</th>
<th>FCVS</th>
<th>AN-L</th>
<th>FCVS-L</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Instantaneous (% TAP)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>1.4</td>
<td>4.9</td>
<td>3.6</td>
<td>2.9</td>
</tr>
<tr>
<td>Mean Memory Pooling</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 1$</td>
<td>16.9</td>
<td>21.4</td>
<td>25.1</td>
<td>24.8</td>
</tr>
<tr>
<td>$m = t$</td>
<td>18.4</td>
<td>30.7</td>
<td>8.5</td>
<td>9.3</td>
</tr>
<tr>
<td>$m = m^*$</td>
<td>21.7</td>
<td>28.8</td>
<td>29.3</td>
<td>30.0</td>
</tr>
<tr>
<td>Max Memory Pooling</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = t$</td>
<td>20.0</td>
<td>27.4</td>
<td>9.0</td>
<td>9.5</td>
</tr>
<tr>
<td>$m = m^*$</td>
<td>21.0</td>
<td>27.5</td>
<td>29.7</td>
<td>30.3</td>
</tr>
<tr>
<td>Memory Welling</td>
<td>22.5</td>
<td>30.5</td>
<td><strong>30.1</strong></td>
<td><strong>30.6</strong></td>
</tr>
<tr>
<td>Max Memory Welling</td>
<td><strong>24.6</strong></td>
<td><strong>35.9</strong></td>
<td>11.0</td>
<td>15.9</td>
</tr>
</tbody>
</table>

- Memory welling outperforms
- Strongest with temporal annotation
- Avg-to-t fails under such conditions
Continuous Retrieval - Results

<table>
<thead>
<tr>
<th>Continuous (% ZP)</th>
<th>AN-L</th>
<th>FCVS-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>1.3</td>
<td>1.1</td>
</tr>
<tr>
<td>Mean Memory Pooling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 1$</td>
<td>21.9</td>
<td>21.6</td>
</tr>
<tr>
<td>$m = t$</td>
<td>5.9</td>
<td>6.3</td>
</tr>
<tr>
<td>$m = m^*$</td>
<td>27.5</td>
<td>27.7</td>
</tr>
<tr>
<td>Max Memory Pooling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = t$</td>
<td>5.9</td>
<td>6.0</td>
</tr>
<tr>
<td>$m = m^*$</td>
<td>27.3</td>
<td>27.5</td>
</tr>
<tr>
<td>Memory Welling</td>
<td>28.3</td>
<td>28.4</td>
</tr>
<tr>
<td>Max Memory Welling</td>
<td>5.6</td>
<td>10.9</td>
</tr>
</tbody>
</table>

→ Max welling copes well with changing content
Analysis - Annotation Type

Frame concepts are objects, which could affect performance on events or scenes

<table>
<thead>
<tr>
<th>Category</th>
<th>MMW</th>
<th>MMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art</td>
<td>20.4</td>
<td>14.7</td>
</tr>
<tr>
<td>Leisure &amp; Tricks</td>
<td>34.0</td>
<td>24.5</td>
</tr>
<tr>
<td>Nature</td>
<td>64.6</td>
<td>55.2</td>
</tr>
<tr>
<td>Travel</td>
<td>31.3</td>
<td>30.0</td>
</tr>
<tr>
<td>Everyday Life</td>
<td>31.2</td>
<td>21.0</td>
</tr>
<tr>
<td>Sports</td>
<td>48.5</td>
<td>32.6</td>
</tr>
<tr>
<td>Beauty &amp; Fashion</td>
<td>24.3</td>
<td>17.1</td>
</tr>
<tr>
<td>Music</td>
<td>35.7</td>
<td>28.8</td>
</tr>
<tr>
<td>DIY</td>
<td>16.9</td>
<td>13.1</td>
</tr>
<tr>
<td>Education &amp; Tech</td>
<td>67.8</td>
<td>51.4</td>
</tr>
<tr>
<td>Cooking &amp; Health</td>
<td>27.7</td>
<td>20.9</td>
</tr>
</tbody>
</table>

→ Strongest performance on objects
→ Suggests need for inclusion of temporal concepts

<table>
<thead>
<tr>
<th>Annotation Type</th>
<th>MMW</th>
<th>MMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Place - Particular location</td>
<td>39.1</td>
<td>26.8</td>
</tr>
<tr>
<td>Object - Thing or creature</td>
<td>67.1</td>
<td>50.0</td>
</tr>
<tr>
<td>Scene - Generic scene setting</td>
<td>39.4</td>
<td>33.3</td>
</tr>
<tr>
<td>Event - Particular occurrence</td>
<td>28.5</td>
<td>21.1</td>
</tr>
<tr>
<td>Activity - Human activities</td>
<td>30.3</td>
<td>22.2</td>
</tr>
</tbody>
</table>
Data Sets

No data set for video stream retrieval
→ Existing data sets can be adapted

**ActivityNet** - 7200 videos, 100 classes, temporal extent

**FCVID Subset** - 239 classes, class diversity

**AN-L** - 30 min videos composed of randomly concatenated short videos
Experimental Setup

Features
- 13k ImageNet concepts, GoogLeNet architecture
- Sampled 2x per second

Setup
- Videos evaluated “concurrently”
- Only present and past frames
- Validation and Test split
- Disjoint classes
Max Memory Welling

Memory Welling can be adapted to traditional video tasks

Max pooling over the welling values \( w(x_t) \)

\[
\text{score}(q, x_t) = \max_{i=0}^{t} s(q)^T w(x_i)
\]

→ Leverage high-confidence temporally local scores to whole video retrieval
Mean and Max Memory Pooling

Memory pooling restricts knowledge to limited temporal window into past

$$MP_{mean}(x_t) = \frac{1}{m} \sum_{i=t-m}^{t} x_i$$
$$MP_{max}(x_t) = \max_{i=t-m}^{t} x_i$$

→ Like temporal sliding window tethered to present
→ Hard memory horizon beyond which all is lost
Computational Expense

...